Jose A González

Professor Mazin Al Hamando

Information Retrieval

9th May 2019

# Using Semantic Networks for Multilingual Information Retrieval

## Abstract

As English becomes increasingly more important in our globalized internet, the retrieval of documents in one language using another also becomes more relevant. While attempts in the past have been done to improve multilingual information retrieval using translators, the use of semantic networks in this area is surprisingly underwhelming. With this project, the author aims to analyze the implementation of lexical databases in multilingual search engines, and provide the advantages and disadvantages of such a take.

While widening the possible translation of a word through this method helps to predicts said translation, it also increases the chance of the query being misinterpreted. Further implementation of semantic networks and linguistic databases may help narrowing down the valid semanticizations, allowing for more accurate results.

## Keywords: multilingual, semantic network, information retrieval, lexical database

## Introduction

As of 2018, English is the most spoken language in the world with 1.121 billion speakers worldwide (Lane). As our world becomes more globalized, the importance of the English language has grown as well, which in turn has lead to a favoritism in optimizing tools for it, often at the expense of improving it for other languages. This issue is no stranger to search engines, and in this particular instance, I am going to tackle the difficulties of finding a work (or its translation) originally in one language, such as English, using another language, such as Spanish.

One of the issues most search engines encounter with this premise, come from the difficulties involved with translation itself. A sentence can be transformed to another language in multiple ways, according to the interpretation and the subject matter itself, creating an unnecessary barrier for works that may already have a translation, but lack a distinct title with the same direct meaning in multiple languages. The aim of this project is to experiment with semantic networks to find new ways of translating queries in information retrieval, hence facilitating the multilingual search of documents.

## Problem Statement

I decided to focus on finding the corresponding articles, rather than on making an efficient and polished web crawler. The main objective of this project is to experiment with the blend of translators and semantic networks, so further research can be based on this approach. My intent is to identify the drawbacks of and advantages of a search widened by the use of semantic association in a multilingual environment.

**Background**

Research on multilingual information retrieval has been done before. Progress has been made in identifying effective strategies for multilingual searches (Aula and Kellar), such as specifying the country in the query, changing languages during the search, and searching for information in English and then looking for its counterpart in another language. While this is useful for user interface and database building, or just as raw data for further research, it’s implementation is more related to improving an already stablished search engine rather than introducing new techniques for multilingual information retrieval.

The use of semantic networks in information retrieval has been proposed before. In the work by Sussna, he stablished that semantics-free, word-based information retrieval either returns irrelevant items when all meanings of a search terms are used, or misses relevant terms when they are only indexed under related terms. The author proceeded to develop a semantic network that would obtain synonyms from the WordNet lexical database with the indexed terms, and train a neural network to known what synonyms to take depending on the context of the document read using SMART (Sussna). This investigation is similar to my take on the project, but doesn’t deal with the complexities of multilingual search.

Finally, a very similar approach to mine was patented in 2001 in Hiroshima, Japan. The patent describes a document searching system for multilingual documents that takes an input, translates it if required, and looks for possible synonyms of the translation before starting the search of the document (Nosohara). However, unlike my approach, Nosohara’s patent only uses synonyms to consider changes in the translation. While this may suffice for internal official documents, it wouldn’t be enough for web searching.

**Solution**

For this particular demonstration written in Java, I focused on seeking the English and Spanish Wikipedia articles for any Stephen King novel or short story queried by the user. I chose this specific author’s bibliography because it is vast, varied, popular, and translated to forty languages (OCLC Online Computer Library Center, Inc. ), meaning that most of his works with an English article, also have a Spanish counterpart.

There are three main complications I could identify when searching for titles translated from English to Spanish, and I introduced a measure for each:

* ***Semantically related terms:***Sometimes titles are changed radically due to context, lack of direct translation, or copyright. An example of this is how the title of the novel “*Firestarter*” was changed to “*Ojos de Fuego*” (Eyes of Fire) in Spanish.
  + *Solution:* A direct translation of elements of the name or of the novel can be used to find its equivalent. The ‘fuego’ in “*Ojos de Fuego*”, or words like ‘pirocinesis’ in the premise, can be translated to their direct meaning (‘fire’ and ‘pyrokinesis’) and used to track back to “*Firestarter*”.
* ***Translation by Association:*** A word may be too common in one language, or maybe the context makes it hard to use a direct translation. In these cases, a term with an ‘associated meaning’ similar to the original one may be used, which means a direct translation or even a related term will not suffice. For example, the short novel “*The Running Man*” was changed in Spanish to “*El Fugitivo*” (The Fugutive).
  + *Solution:* When possible, the original terms must be kept indexed, since translations of articles will often mention the original terms used in their original language.
* ***Terms being translated to a similar equivalent, but not literally:*** While the original sentence and the translation have related meanings, they might not be directly correlated. An example of this is how “*The Shining*” is translated in Spanish as “*El Resplandor*” (The Radiance) instead of “El Brillo”.
  + *Solution:* This is when semantic networks and lexical databases can be useful. By adding semantically-related terms to the indexing, the range is broaden and alternate translations can be found.

Taking these conflicts into account, and the measures taken against them, the project presented by this document works as follow (also illustrated in Figure 1):

1. The user enters their search, with the library JOrtho spellchecking their input (I-net Software).
2. The input is tokenized into terms and replicared into in two lists, one for Spanish and one for English.
3. All stop words are removed from each list, according to the language. The stop words from each language were obtained from a collection. (Google LLC.)
4. A translation for each term of each list is then added alongside it using the Google LanguageApp API (Google LLC.)
5. BabelNet API (Babelnet.org) is used as a hub to query lexical databases in this program.
   * For the English list of terms, all items in it are queried to the Princeton University WordNet (Fellbaum, Brown and et al.)
   * For the Spanish, the terms are sent to the Spanish Multilingual Central Repository (Atserias et al. and Gonzalez-Agirre et al.) through the Open Multilingual Wordnet (Bond and Paik)
6. “Stephen King” is fed into both lists as one single ‘required’ term to make sure articles without the author’s name are not even considered. Anything the user encapsulated with quoting marks is added to the lists too. (Neither process is illustrated in Figure 1).
7. The Web Crawling starts. The bibliography for each language is loaded using the pre-installed Jsoup library. For each list, all documents linked in its corresponding bibliography are downloaded, lowercased and parsed, counting the number of instances of the terms in the list for each article and scoring them accordingly. The score and the link are inserted in a dictionary, then sorted.
8. The ten links with the highest score for each language are shown to the user.

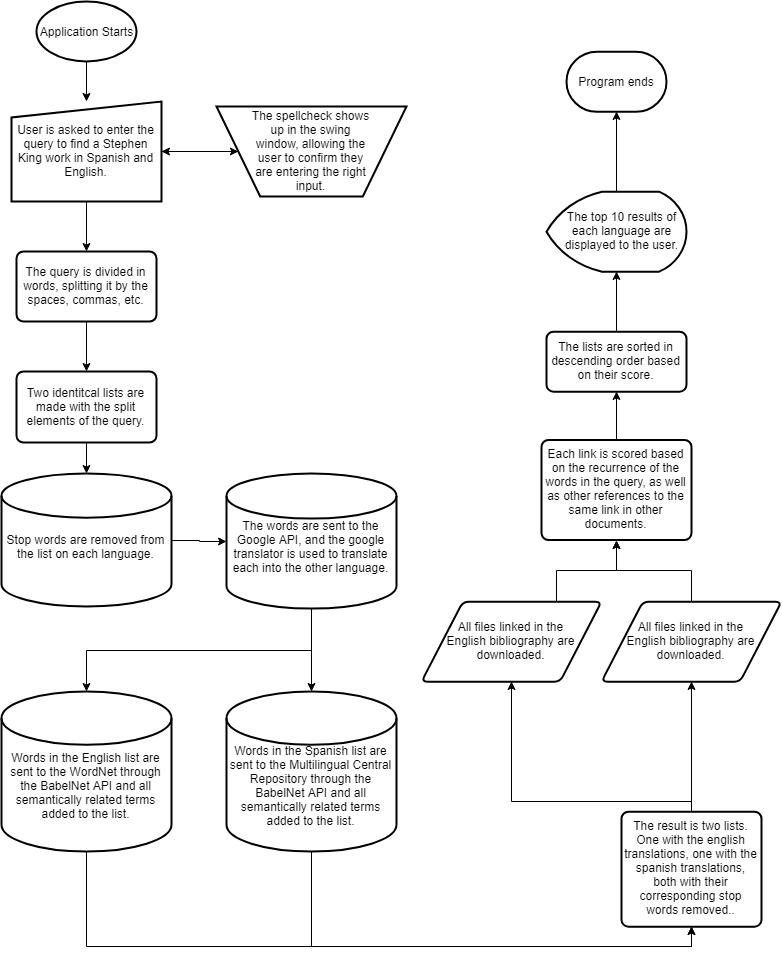
****

Figure - Low Level Design Diagram

## Results

The program was tested first by looking up the titles of the following queries:

* “*The Shining*” produced the best results (Figure 2). Both novels ranked top in their respective list. The lexical database had a connection between ‘shining’ and ‘radiance’, which in turn gave us the best result.

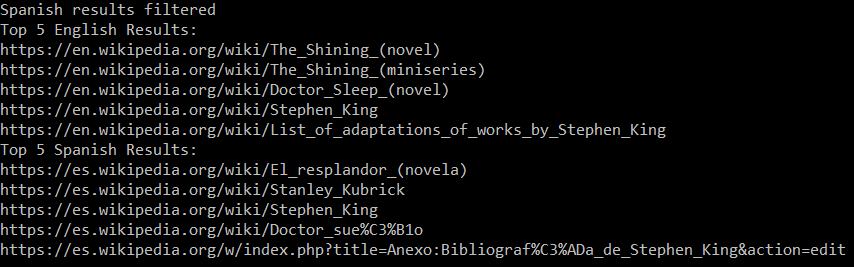


Figure - Results of searching 'The shining'

* “*El Resplandor*” however, performed poorly (Figure 3). The Spanish result was returned as expected, but it seems like the lexical database did not return ‘shining’ as a related term to ‘radiance’, or perhaps the translator differs in its Spanish-to-English translation from its English-to-Spanish one.

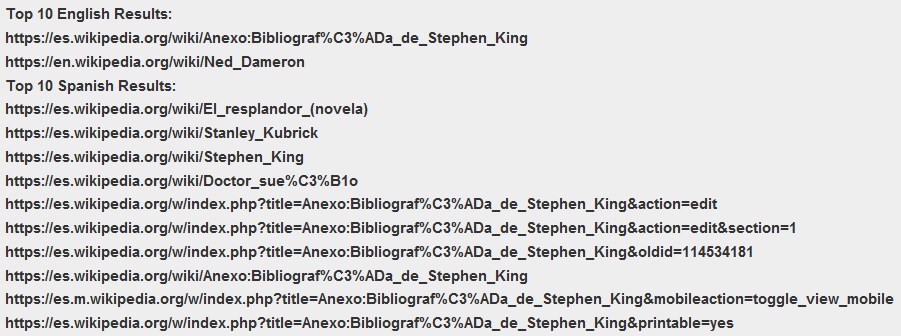


Figure - Results of 'El Resplandor'

* “la zona muerta” (*The Dead Zone*) found both articles related to the novel, but ranked them lower than wider topics with longer articles, and therefore more mentions of the terms added by the semanticization (Figure 4). This result was predictable. The official translation of said title is literal, but its components are also too common.

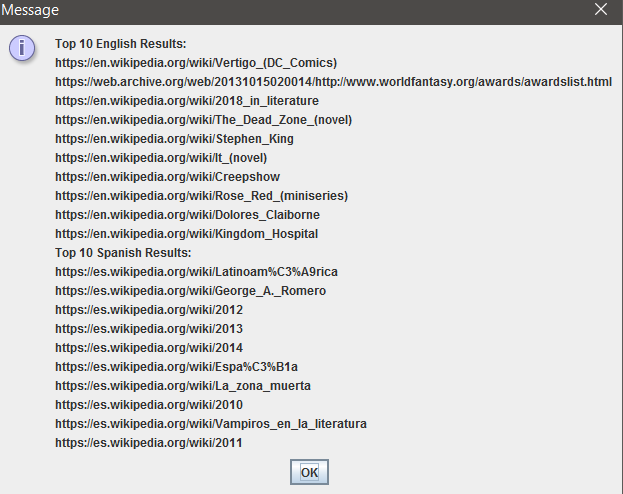


Figure 4 - Results of ‘la zona muerta'

* “fire” was meant to return *Firestarter* and *Ojos de Fuego*. Like with *The Dead Zone*, this one returned the desired results, unsurprising considering they both have the word fire in the title. It’s interesting that the assassination of John F Kennedy ranked higher in the Spanish version, probably because MCR offered ‘disparo’ (gunshot) as a related term to ‘fire’. (Figure 5)

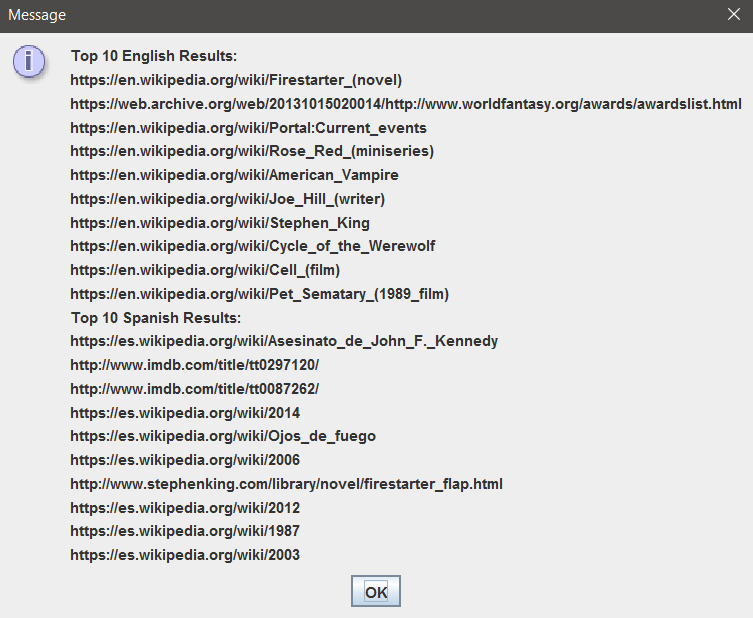


Figure - Results for 'fire'

* “Firestarter”, an even more precise query, still returned our desired results, showing the importance of keeping the original term before and after translation. Interestingly enough, when doing this search we get the IMDB sites ranked higher than the Wikipedia article, probably because the lack of a Spanish IMDB means the Spanish Bibliography still mentions the English website in its citations (Figure 6).

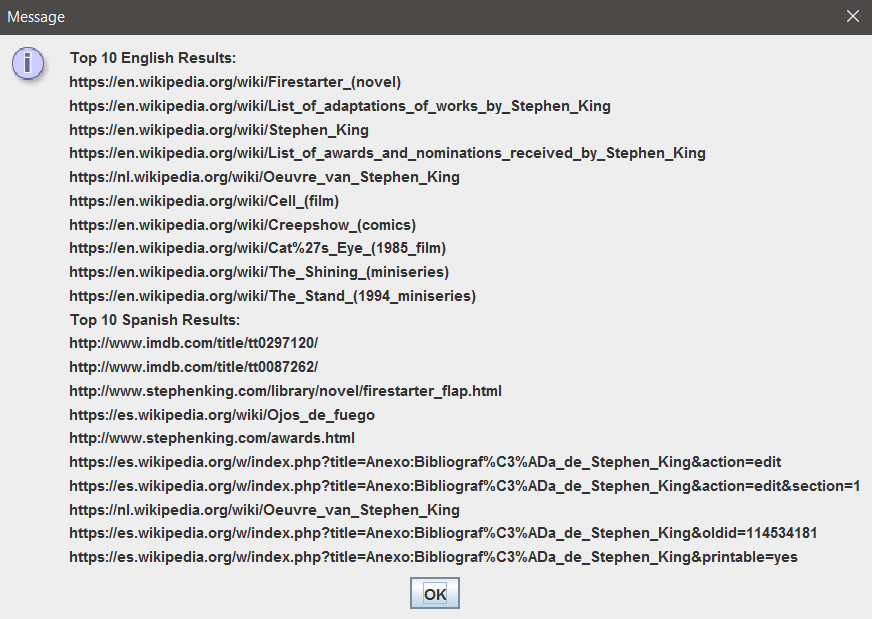


Figure - Results for 'Firestarter'

* “Overlook Hotel”, the place where most of The Shining occurs, successfully returned both results. Probably due to a similar situation as the last query. A very specific term, when kept in the search, will still show up in translations.

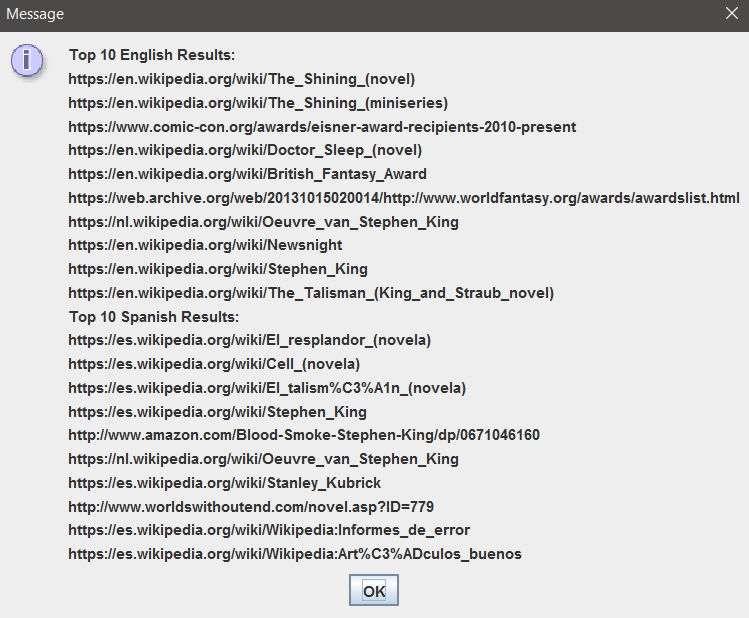


Figure - Results for 'Overlook Hotel'

* “Pennywise”, the name of the scary clown in “*It*” started showing issues with our web crawler. To save time, the depth limit of the web crawler was set to 1. While this works perfectly in the English bibliography (It will automatically redirect to the article for the novel), the Spanish bibliography links to the disambiguation for “It” instead, which means the Spanish article for the novel is not directly accessed from the bibliography. The only direct mention of the monstrous clown apparently can only be found in Stephen King’s article.

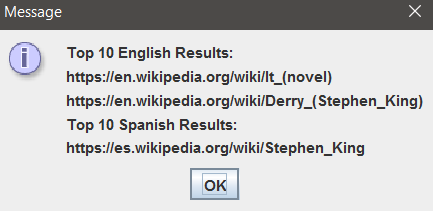


Figure - Results for 'Pennywise'

* “pyrokinetic”, the power to control fire, was queried to retrieve “*Firestarter*” again. Apparently, it couldn’t find the mention of the word in either language, so it settled with the IMDB article cited in the Spanish bibliography. (Figure 9)



Figure - Results for ‘pyrokinetic’

* Finally “El Fugitivo” was queried to get its English counterpart, “The Running Man”. As I expected beforehand, it couldn’t find the English article. This is because ‘fugitive’ is not listed in WordNet in the same category as ‘running man’ or anything of sorts. So the translation offers no aid in finding the proper counterpart. (Figure 10)

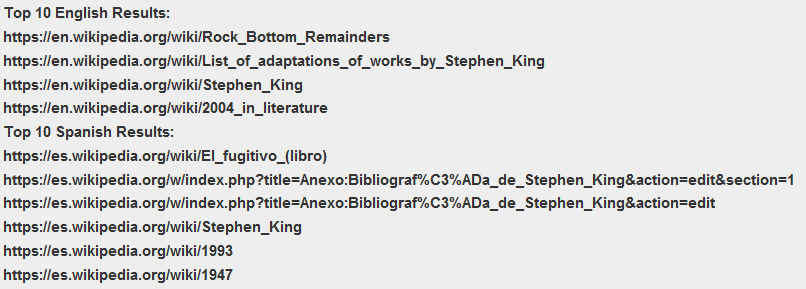


Figure - Results for 'El Fugitivo'

## Conclusion and Future Work

While the introduction of semantic networks helped in the retrieval of documents in a multilingual environment, it did not suffice by itself to accurately predict translations. By widening the possible linguistic interpretation of a term, it offers alternate translations to the same sentence but it also increases the chance of a false positive, particularly when said system is attached to a basic Web Crawler that scores based on word counting. I don’t think, however, that the idea is worthless to pursuit, so I propose the following solutions for further development in this area:

1. ***Create a Context Sensitive Semantic Network***, similar to the one developed by (Sussna). By training a neural network to recognize similitudes in ideas and concepts, additional to the translations themselves, the program will be able to make connections between seemingly separate concepts. For example, this could help close the gap between a ‘fugitive’ who is on the run, and a ‘running man’, or to dismiss the word ‘gunshot’ after searching ‘fire’.
2. ***Building a database from usage*** provides exceptions that may skip over the translation entirely. By studying what results users end up clicking with after entering a query, we can study the relation between languages and construct a reference for future translations.
3. ***Implement these approaches in a better search engine***, the web crawler used in this project served well its purpose for such a small scope, but it handicapped the performance of the experiment by purely basing its score on word iteration, and not frequency, not to mention its huge waiting time between query and result (over 4 minutes).

With these proposed developments I invite further research into the applications of semantic networks and lexical databases in multilingual information retrieval. A more in-depth research into the matter using machine learning to determine context, and a better search engine to score appropriately, may finally solve the many issues of looking for a document in a language that is not your native, helping to communicate our increasingly globalized cultures.

*"I pledge that on all academic work that I submit, I will neither give nor receive unauthorized aid, nor will I present another person's work as my own."*

# References

Atserias et al. and Gonzalez-Agirre et al. *Multilingual Central Repository*. 2012. <http://adimen.si.ehu.es/web/MCR/>.

Aula, Anne and Melanie Kellar. "Multilingual Search Strategies." *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (2009): 3865-3870. Handout.

*Babelnet.org*. n.d. 08 05 2019. <https://babelnet.org/about>.

Bond, Francis and Kyonghee Paik. *Open Multilingual Wordnet*. 2012. 09 May 2019. <http://compling.hss.ntu.edu.sg/omw/>.

Fellbaum, Christiane , Keith Brown and et al. *WordNet, A Lexical Database for English*. Princeton University. 2005. <https://wordnet.princeton.edu/>.

Google LLC. *Google LanguageApp API*. n.d. <https://developers.google.com/apps-script/reference/language/language-app>.

—. *Google Script API*. n.d. <https://www.google.com/script/start/ >.

—. *Stop Word Collection*. November 2011. <http://code.google.com/p/stop-words/>.

Lane, James. "The 10 Most Spoken Languages In The World." 14 11 2018. *Babbel Magazine.* 14 November 2018. <https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world>.

Nosohara, Makifumi. Document Searching System for Multilingual Documents. United States: Patent US 6,212,537 B1. 21 09 1999. <https://patentimages.storage.googleapis.com/e0/21/2f/b50db58f1d615d/US6212537.pdf>.

OCLC Online Computer Library Center, Inc. . *King, Stephen 1947- | WorldCat Identities*. n.d. 09 May 2019. <https://worldcat.org/identities/lccn-n79-63767/>.

Software, I-net. *Jortho Library*. n.d. <https://www.inetsoftware.de/other-products/jortho>.

Sussna, Michael. *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*. 1993. Document. 08 05 2019. <https://sites.ualberta.ca/~golmoham/SW/ontology%20based%20similarity/Word%20sense%20disambiguation%20for%20free-text%20indexing%20using%20a%20massive%20semantic%20network.pdf>.